AD660120

RADC-TR- 67-472
Final Report

AUTOMATIC SECURITY CLASSIFICATION STUDY

Isadore Enger
Guy T. Merriman
Ann L. Bussemey
Travelers Research Center, Incorporated

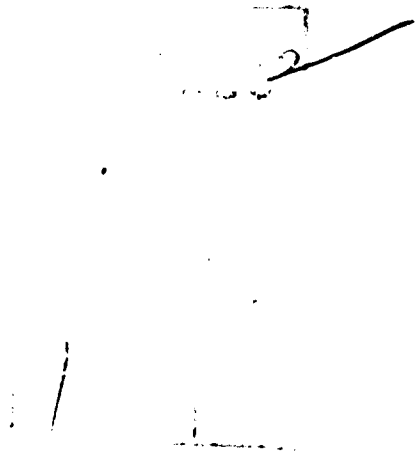
TECHNICAL REPORT NO. RADC-TR-67-472
October 1967

This document has been approved for public release and sale; its distribution is unlimited.

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York

61

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded, by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacturer, use, or sell any patented invention that may in any way be related thereto.



Do not return this copy. Retain or destroy.

AUTOMATIC SECURITY CLASSIFICATION STUDY

Isadore Enger
Guy T. Merriman
Ann L. Bussemey

Travelers Research Center, Incorporated

This document has been approved
for public release and sale; its
distribution is unlimited.

AFLC, GAFB, N.Y., 2 Nov 67-54

FOREWORD

This final report was prepared by Isadore Enger, Guy T. Merriman, and Ann L. Bussemey of Travelers Research Center, Incorporated, 250 Constitution Plaza, Hartford, Connecticut, under Contract F 30 602-67-C-0042, project number 5581, task number 558102. RADC project engineer is Nicholas M. DiFondi (EMIIH).

This technical report has been reviewed by the Foreign Disclosure Policy Office (EMLI) and the Office of Information (EMLS) and is releasable to the Clearinghouse for Federal Scientific and Technical Information.

This report has been reviewed and is approved.

Approved: FRANK J. TOMAINI
Chief, Info Processing Branch
Intel & Info Processing Div.

Approved: JAMES J. DIMEL, Colonel, USAF
Chief, Intel & Info Processing Div.

FOR THE COMMANDER: IRVING J. GABELMAN
Chief, Advanced Studies Group

ABSTRACT

An investigation was made of the feasibility of using computers to assign the proper security classification (unclassified, confidential, secret) to textual material. The words in 998 paragraphs were transformed to computer-usable form. A set of 66 variables was computed for each paragraph by a two-stage process of attaching three scores to a word and then combining the scores in various ways over the words of a paragraph. Several experiments were conducted to validate assumptions involved in the method of scoring the words and the methods for combining the scores. The 66 variables were presented to a statistical technique which made a preferential selection of a small set of effective variables from the large set of 66 variables. The redundant or non-controlling variables were eliminated from subsequent analysis, and an objective system was developed for assigning security classifications using only the selected variables. The system was applied to an inuependent sample of paragraphs and 53.9 percent were correctly classified. It was concluded that the system does exhibit skill. However, the skill is probably too low to consider replacing the present system. Finally, it is concluded that the method for forming variables and the statistical technique, both apparently new to this field, show sufficient promise to merit application to other automatic indexing problems.

TABLE OF CONTENTS

LIST OF TABLES

EVALUATION

The objective of this effort was to determine the feasibility of using a computer algorithm to assign to paragraphs their proper security classification (unclassified, confidential, secret).

Multiple discriminant analysis and regression estimation of event probabilities techniques were used on a dependent sample of paragraphs to develop the algorithm. An independent sample was used to test the algorithm.

Experimental results have shown that the algorithm has predicted the proper security classification at a level much higher than could be attained by chance. However, this level is too low to warrent using the algorithm in place of present classification methods. The degree of compromise of classified information is high as well as the degree of overclassification.

Separate analysis of each security category reveals that in 9 out of 12 experiments the algorithm exhibits difficulty in assigning the proper security classification to confidential material. This indicates that the confidential category does not contain enough specific information to allow the algorithm to distinguish it from the other categories.

This study implies that before automatic security classification can be realized, confidential and secret material should be examined to determine how much overclassification exists. Human classifiers could reveal to the researcher what experience factors are used in classifying textual data. Even if a completely automatic algorithm cannot be developed, it may prove to be sufficiently accurate to aid the human classifier and reduce his task.

Nicholas M. DiFondi

NICHOLAS M. DiFONDI
Technical Evaluator

SECTION I

INTRODUCTION

The assignment of security classifications to military and government publications is a problem because assessing the importance of the contents of a document is judgmental, time-consuming and potentially expensive. Overclassification imposes unnecessary handling restrictions, limits the dissemination of useful information, and impedes the further development of ideas. Underclassification compromises the very values that classification seeks to protect. In the face of the "information explosion," imposing increasing demands on document monitors, other means of classifying textual material are being sought.

The feasibility of using computers was investigated to assign the proper security classification (unclassified, confidential, secret) to textual material. Statistical analysis of the frequency and distribution of words within a paragraph led to a computer-automated procedure of security classification. The procedure was tested for accuracy on independent data by comparing its security assignments with those made subjectively.

SECTION II

PREPARATION OF DATA

The words in 998 paragraphs classified as either unclassified, confidential, or secret were punched onto IBM cards and then placed onto magnetic tape. Described below are criteria for selection of paragraphs, rules for punching data, editing procedures, elimination of function words, and generation of word pairs.

1. Selection of Paragraphs

All documents were chosen from the chemical-biological warfare field. The decision to confine attention to just one field of knowledge is desirable for two reasons: to reduce the number of different words and to obtain a homogeneous sample of paragraphs. It is well-known in linguistic studies that for a given number of words, the number of different words increases rapidly with the number of different fields of knowledge. This smaller number of different words is much easier to handle both from a statistical viewpoint (e.g., the sampling variabilities of frequently occurring words are less than those of infrequent ones) and from a data processing viewpoint (less data are easier to process).

It is even more important to have a homogeneous sample of paragraphs. The overall objective of this study is to discriminate among paragraphs with different security classifications. This is best accomplished by minimizing paragraph differences caused by any circumstance other than security classification for these can only obscure the differences due to security classification. Specifically, different fields of knowledge imply different paragraph content which might overwhelm the differences due to security classification. The chemical-biological warfare field was chosen because a number of documents were readily available from previous studies.

It would be desirable to select all paragraphs from one document which would provide a homogeneous sample of paragraphs. As this was not possible, a number of documents were chosen. Documents were selected if each individual paragraph was classified separately and if the document contained all three types of paragraphs--unclassified, confidential, and secret. Forty-two documents were chosen.

Within each document the selection of paragraphs was governed by the following criteria:

(a) length between 50-200 words, in some cases small consecutive paragraphs with the same classification were combined to achieve the minimum length;

(b) none, or very few, formulas, equations, or other non-word characteristics; and

(c) as far as possible paragraphs with different classifications were alternated (i.e., C, S, U, S, C, U, etc.).

Nine hundred ninety-eight paragraphs were chosen: 341 unclassified, 338 confidential, and 319 secret.

2. Card-Punching Rules

There are available a number of sets of rules for card-punching textual material. However, these were devised for the purpose of semantic and syntactic analysis and were deemed to be too elaborate for this study. Therefore, a set of rules was devised and kept as simple as possible to minimize the number of errors.

At the beginning of each paragraph a header card was punched containing only the number of the paragraph and its security classification. Punching of words started in column one of the next card, continued through column 70, then

onto column one of the next card through column 70, etc. One word could overlap two cards. Characters not appearing on the key-punch keyboard were ignored. Two periods were used at the end of a sentence. Numbers referring to footnotes were not punched but exponents of variables were punched.

3. Definition of "Word," "Word-Pair," and "Function Word"

3.1 Word

All information from a single paragraph was treated as one long serial string of characters. It will be recalled from Section 2.2 that characters that do not appear on the IBM key-punch were not punched. The characters that do appear were separated into two kinds--permissible and non-permissible. The 18 non-permissible characters are:

<	Less-than Sign	,	Comma
(	Left Parenthesis	%	Percent
|	Vertical Bar, Logical OR	—	Underscore
&	Ampersand	>	Greater-than Sign
!	Exclamation Point	:	Colon
*	Asterisk	#	Number Sign
)	Right Parenthesis	@	At Sign
;	Semicolon	'	Prime, Apostrophe
¬	Logical NOT	"	Quotation Marks

A "word" is defined as a string of successive characters bounded by either a blank, a non-permissible character, or an end-of-sentence mark.

3.2 Word-Pair

A word-pair is defined here as two consecutive words in the same sentence. For example, in the sentence GEORGE WASHINGTON CROSSED THE DELAWARE.. the pairs

are GEORGE WASHINGTON, WASHINGTON CROSSED, and CROSSED DELAWARE. Note that THE is eliminated because it is a function word.

3.3 Function Words

Function words include those which are traditionally called articles, prepositions, pronouns, conjunctions and auxiliary verbs, plus certain irregular forms. The list below defines the function words of this study. (The reason for truncation to six letters will be given later.)

-	ARE	DOES	PAST	ALONE	UNTIL	EITHER	NEITHE	THINGS
A	BUT	DONE	PLUS	ALONG	WASNT	ELSEWH	NEVERT	THOUGH
I	CAN	DONT	REAL	AMONG	WHERE	ENOUGH	NOBODY	THROUG
AM	DID	DOWD	SAME	APART	WHICH	EVERMO	NOTHIN	TOGETH
AN	ETC	EACH	SELF	ASIDE	WHILE	EVERYO	NOWADA	TOWARD
AS	FEW	ELSE	SOME	BEING	WHOSE	EVERYT	NOWHER	UNDERN
AT	FOR	EVEN	SUCH	BELOW	WOULD	EVERYW	OFTENT	UNDOIN
BE	GET	EVER	THAN	COULD	ACROSS	EXCEPT	OTHERS	UNLESS
BY	GOT	FROM	THAT	DOING	AGAINS	FAIRLY	OTHERW	WHATEV
DO	HAD	GETS	THEM	EVERY	ALREAD	FARTHE	OURSEL	WHENEV
HE	HAS	HAVE	THEN	LATER	ALMOST	FOREGO	OUTSID	WHEREA
IF	HOW	HERE	THEY	LEAST	ALTHOU	FOREVE	OUTWAR	WHEREF
IN	ITS	INTO	THIS	MIGHT	ALWAYS	FORWAR	OVERMU	WHEREI
IS	MAY	JUST	THUS	NEVER	AMOUNT	FURTHE	PERHAP	WHEREV
IT	NOW	KEEP	UNTO	OFTEN	ANOTHE	HARDLY	PLEASE	WHETHE
ME	OUR	KEPT	UPON	OTHER	ANYBOD	HAVING	PRETTY	WITHIN
MY	OWN	LESS	VERY	OUGHT	ANYONE	HEIGHT	RATHER	WITHOU
NO	THE	LEST	WELL	QUITE	ANYTHI	HENCEF	REALLY	YOURSE
OF	TOO	MANY	WERE	RIGHT	ANYWHE	HEREIN	SEVERA	
OH	WAS	MINE	WHAT	SHALL	AROUND	HITHER	SHOULD	
ON	WAY	MORE	WHEN	SHALT	AWFULL	HOWEVE	SOMEBO	
OR	WHO	MOST	WHOM	SINCE	AWHILE	INDEED	SOMEDA	
SO	WHY	MUCH	WILL	STILL	BACKWA	INSTEA	SOMETH	
TO	YES	MUST	WILT	THEIR	BECAUS	INWARD	SOMETI	
UP	YET	NEXT	WITH	THERE	BEFORE	ITSELF	SOMEWH	
US	YOU	NONE	YOUR	THESE	BEHIND	LIKEWI	THEIRS	
WE	ALSO	ONES	ABOUT	THING	BETWEE	MIDDLE	THEMSE	
ALL	AWAY	ONLY	ABOVE	THOSE	BEYOND	MIGHTY	THEREA	
AND	BEEN	ONTO	AFTER	TRULY	CANNOT	MOREOV	THEREF	
ANY	BOTH	OVER	AGAIN	UNDER	DURING	MYSELF	THEREW	

4. Editing Procedures

A three-stage procedure was followed to correct errors.

Proofreading. The cards were listed and the listings were proofread. Errors were corrected by punching new cards.

Misspellings. The cards were placed onto magnetic tape and the words were extracted using the definition of Section 2.2.1. There were 112,774 words in all. These were alphabetized by the computer and all words appearing only once or twice were printed by the computer. The rationale here is that the same misspelling will not occur more than twice. The printouts were proofread and misspellings were corrected by repunching cards.

Hyphens. Depending upon the author, the same word may or may not be hyphenated, e.g., anti-aircraft, antiaircraft. These words presented a special problem and were examined carefully. A correction card eliminating the hyphen was punched for those hyphenated words which we considered to be two words. The hyphens within the remaining hyphenated words were eliminated by "squeezing-up" those words. After correction for hyphens the misspelling edit was repeated.

5. Form of Data

Function words were eliminated from the edited data because it was felt that they would not contribute to discriminating among unclassified, confidential, or secret paragraphs. The function words constituted some 42 percent of the total number of words and their elimination saved considerable computer time.

All words were truncated to six letters. In word studies it is desirable--and it is common practice--to define two words with the same root but different endings as the same word. However, the rules for so doing are quite elaborate

and it was not deemed worthwhile to write the complicated computer programs necessary to apply such rules. The truncation is a simple expedient to combine words with the same root. It does occasionally result in combining two words with different roots and it does not combine words of less than six letters. Nevertheless, it is a rather effective substitute for the more accurate, but much more complicated, procedures now in use.

All words and word-pairs were then placed onto a magnetic tape for subsequent processing.

SECTION III

STATISTICAL METHODOLOGY

Two statistical procedures were used to develop objective methods for assigning security classifications to paragraphs. The techniques are (1) stepwise linear regression, and (2) regression estimation of event probabilities (REEP). The techniques have been described in some detail in other publications: stepwise regression in [5] and REEP in [4]. Detailed descriptions of applications of these techniques are available in [7] and [3]. In this section we simply describe the procedures; their application is considered in later sections.

The 998 paragraphs were separated into two samples by extracting every third paragraph. The statistical procedures are applied to the larger--or developmental--sample, and the resulting methods for assigning security classifications were tested for accuracy on the independent sample.

In both techniques, a stipulated variable--security classification--called the predictand is the object of estimation. The variables used to make the estimation of the predictand are called predictors. Both techniques begin with computation of a "predictor-predictand" matrix as in Table I.

The general entry, X_{nm}, in Table I is the value of the m-th predictor in the n-th paragraph. The formation of predictors is discussed in detail in Section 4. The predictand variables, the Y's, are dummy variables--i.e., variables which can take on only the values zero or one. For example, Y_{nU} takes on a value of one if paragraph n is unclassified and $Y_{nC} = Y_{nS} =$ zero; Y_{nC} is one if a paragraph is confidential and $Y_{nU} = Y_{nS} =$ zero; and similarly for Y_{nS} and secret. N is the number of paragraphs in the developmental sample.

TABLE I

PREDICTOR-PREDICTAND MATRIX

Paragraph Number	Predictor Number						Predictand		
	1	2	...	m	...	M	U	C	S
1	X_{11}	X_{12}	...	X_{1m}	...	X_{1M}	Y_{1U}	Y_{1C}	Y_{1S}
2	X_{21}	X_{22}	...	X_{2m}	...	X_{2M}	Y_{2U}	Y_{2C}	Y_{2S}
⋮									
n	X_{n1}	X_{n2}	...	X_{nm}	...	X_{nM}	Y_{nU}	Y_{nC}	Y_{nS}
⋮									
N	X_{N1}	X_{N2}	...	X_{Nm}	...	X_{NM}	Y_{NU}	Y_{NC}	Y_{NS}

The number of plausible predictors that might serve to assign security classifications to paragraphs is very large, if not virtually unlimited. For example, X_{nm} could be a frequency count of the number of times that word m occurred in paragraph n, in which case the number of predictors M would be equal to the number of different words in all N paragraphs. This situation imposes the practical necessity of selecting a manageable number of predictors. The statistical techniques, therefore, include provisions for the preferential selection of effective predictors from a very large set of possible choices for use in regression or REEP. Substantial previous experimentation comparing performance on independent data of estimating functions using large numbers of predictors with those using selectively chosen subsets of such variables has shown, as a rule, that whatever predictability resides in a large set is almost wholly contained in the much smaller subset. The objective selection of such a small subset is termed screening. After screening, the redundant or non-controlling predictors are eliminated from subsequent analyses, and a system for assigning security classifications to paragraphs is developed using only the selected predictors.

6. Stepwise Linear Regression

In multiple regression, a predictand Y is expressed as a linear function of a number M of predictor variables X_m (m=1,2,...,M).

$$\hat{Y} = A_0 + A_1X_1 + A_2X_2 + \ldots + A_MX_M \qquad \text{(III-1)}$$

where the coefficients A_m (m=0,1,...,M) are determined by least squares. Y can be an estimate of any one of the three Y's of Table I, i.e., the stepwise technique is applied three times.

As noted above, if M is large, screening is desirable. To select the first predictor, the simple linear correlation coefficient is computed between the predictand Y and each of the entire set of M predictors. The predictor giving the best coefficient (i.e., highest in absolute value) is selected as the first predictor. Next, the partial correlation coefficient between each of the remaining predictors and the predictand (holding the first selected predictor constant) are examined, and the predictor associated with the best coefficient is then selected as the second predictor. Additional predictors are selected in a similar manner. At each step partial correlations are computed between the predictand and each of the remaining predictors while holding constant the previously selected predictors. The predictor associated with the best partial correlation is selected. This is equivalent to selecting that predictor which adds the most independent predictive information to the previously selected predictors. At each step a test is made to see if the new predictor selected adds a satisfactory amount of additional information. When the test fails selection is halted. A multiple regression is then computed between the variable to be predicted and the small set of selected predictors.

The stepwise regression technique will result in three equations--one for each security classification. In applying the equations to the test sample of paragraphs, the largest $\hat{Y}$ gives the security classification assignment.

7. Regression Estimation of Event Probabilities (REEP)

The REEP technique is much like stepwise regression but differs in two important respects: (1) the use of dummy predictors exclusively, and (2) the simultaneous consideration of all three predictand dummy variables--the three Y's of Table I--rather than piecemeal consideration.

The first step in the REEP procedure is to transform each predictor variable to a set of "dummy variables." A dummy variable is a variable which can take on only two values, zero or one. An example illustrates the procedure for continuous variates. If X is a continuous variable, it can be divided into G ranges by specifying G-1 class limits, X_1, X_2, ..., where

$$-\infty < X \leq X_1,\ X_1 < X \leq X_2,\ \ldots,\ X_{G-1} < X < +\infty \quad .$$

A set of G dummy variables is generated by finding the range which encloses a specific X-value and assigning a one (1) to the proper dummy variable and zero (0) to the remaining G-1 dummy variables. This procedure is repeated for all N values of X. Qualitative variables are transformed into dummy predictors in a manner similar to that indicated in Table I for the predictand "security classification." This permits qualitative variables to be incorporated into the REEP procedure in a natural and easy manner.

The selection of predictors is made by first computing the simple linear correlation coefficient between each dummy predictor and each dummy predictand --or 3 times M coefficients in all. The highest coefficient of this entire set gives the first predictor. Next, the partial correlation coefficients between <u>each</u> of the remaining predictors and <u>each</u> of the dummy predictands are examined and the highest one gives the second predictor. Additional predictors are selected in a similar manner. Finally, three regression equations are computed, one between each of the three predictand variables and the final set of selected predictors. Application of the equations to the test sample of paragraphs will give the probabilities that a paragraph belongs to each of the three security classifications.

SECTION IV

FORMATION OF PREDICTOR VARIABLES

The basic assumption underlying this study is that the words of a paragraph contain information for determining its security classification. To extract such information, 66 predictor variables which potentially contain information were computed. These were arranged in a matrix as indicated in Table I and the two statistical techniques were applied to select in an objective manner those predictor variables which actually contain information for discriminating among the three types of paragraphs.

Three decisions discussed previously impact on the formation of predictor variables: (a) function words were eliminated, (b) word-pairs are considered as words, and (c) the developmental sample of paragraphs is used for generating the matrix.

Predictor variables were formed by a two-stage process of first attaching three scores to each word and then combining the scores over the words in a paragraph. The combination was done in a variety of ways resulting in a number of predictor variables. In this section we describe the method of forming the predictor variables. Their use is discussed in subsequent sections.

8. Word Scores

Three scores were assigned to each word. As a first step in devising the scores, it was assumed that it is simply the appearance or non-appearance of a word in a paragraph, rather than the number of times it appears, which serves to determine the security classification of the paragraph. Consequently, a count was obtained of the number of different paragraphs in which each word appears. This number was designated as N_j for the j-th word. Some of these N_j paragraphs were unclassified, some confidential, and some secret. These three quantities were denoted as N_{ju}, N_{jc}, and N_{js}, where

$$N_{ju} + N_{jc} + N_{js} = N_j \quad . \tag{IV-1}$$

The convention was adopted that an individual word will be counted each time it appears whether alone or as part of a word-pair; e.g., if word j appears eight times alone and four times as part of a word-pair, denoted as word k, then $N_j = 12$ and $N_k = 4$.

If it is assumed that word j does not offer any information whatsoever about the security classification of paragraphs, then the number of different unclassified paragraphs in which word j should appear is proportional to the total number of unclassified paragraphs. (Such an assumption is termed the null hypothesis in statistical decision theory.) Similarly, the number of different confidential and secret paragraphs in which word j should appear is proportional to the total number of such paragraphs. Mathematically,

$$\begin{aligned} E_{ju} &= N_j(K_u/K) \\ E_{jc} &= N_j(K_c/K) \\ E_{js} &= N_j(K_s/K) \quad , \end{aligned} \tag{IV-2}$$

where E is termed the expected number of paragraphs, K is the total number of paragraphs in the developmental sample, and K_u, K_c, and K_s are, respectively, the number of unclassified, confidential, and secret paragraphs in this same sample.

Three scores were attached to word j:

$$\begin{aligned} S_{ju} &= (N_{ju} - E_{ju})/(E_{ju})^{1/2} \\ S_{jc} &= (N_{jc} - E_{jc})/(E_{jc})^{1/2} \\ S_{js} &= (N_{js} - E_{js})/(E_{js})^{1/2} \end{aligned} \tag{IV-3}$$

Intuitively, the scores appear reasonable. Their numerators are deviations of actual values from expected values. Thus, if N_{js} exceeds E_{js} by a large amount then word j appears in secret paragraphs much more often than expected by chance. Such a word should be useful for distinguishing between secret and other types of paragraphs. The denominators are factors which take into account the number of paragraphs in which word j appears. This is required since a large value of $(N_{js} - E_{js})$ does not mean the same thing if word j appears in 50 different paragraphs ($N_{js} = 50$), as it does if word j appears in only four paragraphs ($N_{js} = 4$).

Statistically, each S-value is a "chi variate" which approximates a unit normal deviate as N_j increases. Such approximations have desirable properties; in particular, their sum is a meaningful quantity. (An example of non-meaningful quantities is the sum of the numerators alone or the sum of ratios such as N_{js}/N_j.) Considerable effort has been devoted in the field of statistical theory to the minimum value that E can assume for S to have the desirable properties. For the case considered here, where E_{ju}, E_{jc}, and E_{js} are approximately equal because $K_u \cong K_c \cong K_s$, it has been found that $E \geq 1.0$ is satisfactory. Therefore, only those words for which N_j is at least <u>four</u> (4) have scores attached to them.

9. Combining of Scores

There were three different combination methods, and each method resulted in a number of predictor variables.

9.1 Means

The scores were summed in various ways over the usable words of a paragraph. A <u>usable</u> word has three characteristics: (a) it is not a function word, (b) it appears in at least four different paragraphs in the developmental sample of

data, and (c) it has not appeared previously in the same paragraph. Characteristic (c) is necessary to assure that a word appearing more than once in a paragraph will have its scores summed only once. To account for the varying number of usable words in the paragraphs, each sum is divided by the total number of usable words, denoted by L_i for the i-th paragraph.

The first three predictors are the arithmetic means of the three scores over all usable words of the paragraph:

$$\begin{aligned} MU100_i &= (\Sigma S_{ju})/L_i \\ MC100_i &= (\Sigma S_{jc})/L_i \\ MS100_i &= (\Sigma S_{js})/L_i \end{aligned} \qquad \text{(IV-4)}$$

(The reason for the notation is given later.)

Other means were also computed. The rationale for such means can be illustrated by an example. Consider a paragraph with a few large positive values of S_s but many small negative values of S_s. The sum, $MS100_i$, could easily be negative. Yet, logically, it is quite conceivable that there be just a few strong words which make a paragraph secret and the remainder of the words in the paragraph may not be important. To take account of such possibilities, additional means are computed:

$$\begin{aligned} MU\text{--}_i &= (\Sigma \Delta^{(k)}\ S_{ju})/L_i \\ MC\text{--}_i &= (\Sigma \Delta^{(k)}\ S_{jc})/L_i \\ MS\text{--}_i &= (\Sigma \Delta^{(k)}\ S_{js})/L_i \end{aligned} \qquad \text{(IV-5)}$$

$$\text{where } \Delta^{(k)} = \begin{cases} 1 \text{ if } S_{ja} > k \\ 0 \text{ if } S_{ja} \leq k \end{cases}$$

and a can be u, c, or s. Equations (IV-5) state that the scores of the words of paragraph i are summed only for those cases for which the score exceeds k. For example, with k = 0 only positive values of S are summed.

It is not desirable to assume which value of k would result in predictor variables containing the maximum amount of information for discriminating among the three types of paragraphs. Therefore, k was set equal to a number of different values. Insertion of each k in equations (IV-5) results in three predictor variables.

To develop a reasonable set of k-values, advantage was taken of the fact that the scores--the S-values--are approximately normally distributed. Cox [2] and Bryan and Southan [1] have developed a method for the optimum subdivision of a normally distributed variable. Their method was applied as follows: The S_u scores for all words in all paragraphs constitute a normally distributed variable. Similarly, S_c is a normal variate and so is S_s. The three variables are treated separately. A variable is arrayed from lowest to highest. The array is examined to choose class limits which subdivide the variable in the manner indicated in Table II. The class limits are the k-values.

TABLE II

SAMPLE ARRAY TO CHOOSE CLASS LIMITS

Class Limit	Percentage of Scores Higher than Class Limit
k_1	98
k_2	91
k_3	80
k_4	66
k_5	50
k_6	34
k_7	20
k_8	9
k_9	2

Use of the k-values obtained in this way resulted in 27 variables, 9 for each of the three types of scores. The notation of the variables is $MU98_i$, $MC98_i$, $MS98_i$ when k_1 is used in equations (4-5); $MU91_i$, $MC91_i$, $MS91_i$ when k_2 is used, etc.

9.2 Frequencies

Frequency predictor variables are defined in precisely the same manner as Mean variables except that the S-values are counted instead of being summed. Specifically,

$$FU\text{--}_i = (\Sigma\Delta_{ju}^{(k)})/L_i$$

$$FC\text{--}_i = (\Sigma\Delta_{ju}^{(k)})/L_i$$

$$FS\text{--}_i = (\Sigma\Delta_{ju}^{(k)})/L_i$$

where, as before,

$$\Delta_{ju}^{(k)} = \begin{cases} 1 \text{ if } S_{ja} > k \\ 0 \text{ if } S_{ja} \leq k \end{cases}$$

and the summations are made over the usable words of paragraph i.

Twenty-seven predictors were computed by using the 9 k-values listed in Table II: $FU98_i$, $FC98_i$, $FS98_i$, ..., $FU02_i$, $FC02_i$, $FS02_i$.

9.3 Highest Value Sums

There are nine such predictor variables defined as follows:

$HU1_i$ = largest S_u value in paragraph i.

$HU3_i$ = sum of three largest S_u values in paragraph i.

$HU5_i$ = sum of five largest S_u values in paragraph i.

Similar definitions hold for HC- and HS-.

9.4 Summary

The 66 predictors are listed in Table III.

TABLE III

PREDICTOR VARIABLES

MU100	MC100	MS100
MU98	MC98	MS98
MU91	MC91	MS91
MU80	MC80	MS80
MU66	MC66	MS66
MU50	MC50	MS50
MU34	MC34	MS34
MU20	MC20	MS20
MU09	MC09	MS09
MU02	MC02	MS02
FU98	FC98	FS98
FU91	FC91	FS91
FU80	FC80	FS80
FU66	FC66	FS66
FU50	FC50	FS50
FU34	FC34	FS34
FU20	FC20	FS20
FU09	FC09	FS09
FU02	FC02	FS02
HU1	HC1	HS1
HU2	HC2	HS2
HU3	HC3	HS3

SECTION V

EXPERIMENTS

Four experiments, Sections 1 through 13, were completed to test some assumptions made in developing the word-scores and their combination. The paragraphs were separated into two samples--a dependent sample for developing computer methods of assigning security classifications to paragraphs, and an independent sample for testing the accuracy of the methods. An experiment (Section 14) was made to study the optimum size of the dependent sample. In conducting this experiment, it was noted that there was quite a loss in accuracy from the dependent to the independent samples. A procedure (see Section 15) was devised and tested to reduce this loss in accuracy. One of the two statistical techniques was used to develop an automated-computer method for assigning security classifications (see Sections 16 and 17).

10. Individual Words as Predictors

In previous work in development of automatic indexing procedures (e.g., [8]) individual words are used as predictors instead of being scored and then collectively combined. For example, the presence or absence of a specified word in a paragraph could constitute a dummy variable predictor by assigning a one for presence and a zero for absence. Alternatively, the predictor could be the number of times that a word appears in a paragraph. A full-scale investigation of the relative efficacy of such predictors as compared to score-type predictors is beyond the scope of this study. However, some useful information could be obtained in a simple and straightforward manner.

There were 1741 different individual words--exclusive of function words--and 643 different word-pairs appearing in four or more paragraphs. For each of these 2384 "words," frequency counts were made of the total number of paragraphs

the word appeared in and the number of such paragraphs which were unclassified, confidential, and secret. These quantities are the N_j, N_{ju}, N_{jc}, and N_{js} defined in Section 8.

A detailed but entirely subjective examination of these frequencies was made. Many of the words appeared infrequently. Such words cannot be expected to be good predictors simply because most paragraphs will not contain them. Approximately 150 words appearing in 50 or more paragraphs--the frequently occurring words--did not appear to occur more frequently in one type of paragraph than in another (i.e., N_{ju}, N_{jc}, N_{js} did not differ radically). Such words are not useful predictors by themselves.

It was concluded that summing of scores over all usable words of a paragraph would tend to accumulate the small amounts of information in each word and, therefore, would be more likely to prove to be a good prediction method. However, the issue is not closed; and it is recommended that a test be made to determine whether individual word predictors add any information to score-type predictors.

11. Multiple Word Occurrence in a Paragraph

The first occurrence only of a word in a paragraph was used to compute predictor variables. (See Section 9.1 for definition of a usable word.) This assumes that subsequent occurrences contribute no information for assigning a security classification to that paragraph. An experiment was performed to test this assumption.

The test was made by computing a small set of predictor variables using first occurrences only and another similar set using all occurrences. The two sets were then used to assign security classification to paragraphs and the assignments were verified by comparison with the actual classifications.

Three predictor variables MU100, MC100, MS100 were obtained by summing scores for usable words--i.e., first occurrences only are included in the sums. (See equations (IV-4) for definitions of the variables.) Three new predictor variables were computed in a similar manner except that scores for all occurrences of a word were included in the sums. The six variables were computed for each of the 998 paragraphs.

Using the 666 paragraphs of the dependent sample, mean values of each variable were computed. Deviations from mean values were then computed for all 998 paragraphs. To illustrate the computations, consider the variable MU100 of which 998 values were computed by equations (IV-4). The mean is

$$\overline{MU100} = (\Sigma\ MU100_i)/666 \quad , \tag{V-1}$$

where the summation is over all paragraphs in the dependent sample. The deviations are

$$mu100_i = MU100_i - \overline{MU100}; \quad i=1,2,\ldots,998 \tag{V-2}$$

The largest value of $mu100_i$, $mc100_i$, and $ms100_i$ determines the assignment of a security classification to paragraph i. Assignments were made in this way to all 998 paragraphs. Verification of the assignments on the dependent and independent samples for the two sets of variables is shown in Table IV.

The accuracy of the assignments was measured by the proportion of "hits," which is the sum of the elements on the main diagonal divided by the total number of assignments. On the dependent sample, there were 90.3 percent hits for the single-occurrence assignments and 86.6 percent for the assignments made by the multiple-occurrence variables. On the independent sample, the corresponding figures are 51.2 percent versus 49.6 percent. Thus, on both

TABLE IV

COMPARISON OF SINGLE VS. MULTIPLE WORD OCCURRENCE

Dependent Sample

(a) Single Occurrence

Assigned \ Actual	U	C	S	Total
U	199	16	13	228
C	7	202	5	214
S	5	18	201	224
Total	211	236	219	666

$$\% \text{ Hits} = \frac{602}{666} = 90.3$$

(b) Multiple Occurrence

Assigned \ Actual	U	C	S	Total
U	187	10	10	207
C	13	184	3	200
S	11	42	206	259
Total	211	236	219	666

$$\% \text{ Hits} = \frac{577}{666} = 86.6$$

Independent Sample

(c) Single Occurrence

Assigned \ Actual	U	C	S	Total
U	72	35	21	128
C	31	40	20	91
S	28	27	58	113
Total	131	102	99	332

$$\% \text{ Hits} = \frac{170}{332} = 51.2$$

(d) Multiple Occurrence

Assigned \ Actual	U	C	S	Total
U	62	27	14	103
C	31	37	19	87
S	38	38	66	142
Total	131	102	99	332

$$\% \text{ Hits} = \frac{165}{332} = 49.6$$

samples the single-occurrence type variables resulted in better assignments. It is concluded, therefore, that subsequent occurrences of a word in a paragraph do not contribute any information for assigning security classifications.

12. Word-Pairs

An experiment was done to determine whether word-pairs contribute discriminating information. The procedure was quite similar to the one just described. Two small sets of variables, one with and one without word-pairs, were used to assign security classifications to paragraphs. The assignments were then verified and compared. The variables with word-pairs are the mu100, mc100, and ms100 defined in the previous section. A new set of three variables was obtained by summing scores over single words of a paragraph, not including word-pairs. As before, deviations from means were computed, and the highest of the three deviations for a paragraph governed the security assignment.

The assignments were verified on the independent sample only and are presented in Table V.. The percentage of hits is 50.9 without word-pairs and 51.2 with word-pairs. Although the rise is rather small, it is concluded that word-pairs do contribute a small amount of discriminating information over and above that contributed by single words.

13. Four or More Paragraphs

Predictor variables are obtained by considering the scores of usable words only (see Sections 9.1 and 9.2). One characteristic of a usable word is that it must have appeared in at least four paragraphs. Four was chosen because it is the minimum number necessary for S, defined by Equations (4-3), to have certain desirable statistical features. A test was made to see if requiring more than four paragraphs would result in better predictors.

TABLE V

TEST OF USEFULNESS OF WORD-PAIRS

(a) Word-Pairs

Assigned \ Actual	U	C	S	Total
U	72	35	21	128
C	31	40	20	91
S	28	27	58	113
Total	131	102	99	332

$$\% \text{ Hits} = \frac{170}{332} = 51.2$$

(b) No Word-Pairs

Assigned \ Actual	U	C	S	Total
U	70	37	21	128
C	32	38	17	87
S	29	27	61	117
Total	131	102	99	332

$$\% \text{ Hits} = \frac{169}{332} = 50.9$$

Again the procedure was similar to that described in Section 11. Three small sets of variables were obtained: one using the four-paragraph criterion, a second using 10 paragraphs as the criterion, and a third with 15 paragraphs. For each criterion three variables were computed as before. The four-paragraph variables are the same mul00, mc100 and ms100 used in Sections 11 and 12. Two new sets of variables were obtained by redefining a usable word to include (a) only those words appearing in 10 or more paragraphs, and (b) only words appearing in 15 or more paragraphs. New MU100, MS100 and MC100 variables were computed, and deviations from their means were used to assign security classifications. The verifications on the independent sample of data are presented in Table VI. The percentage of hits was 51.2 for the four-paragraph criterion, 51.2 for 10 paragraphs and 50.6 for 15 paragraphs. Since neither of the other criteria gave results better than the four-paragraph criterion the latter was retained.

14. Size of Sample

How many paragraphs are required in the developmental sample to develop an objective method of assigning classifications? It is well-known that use of large samples would result in more accurate assignments. On the other hand, the original subjective assignment of security classifications to a developmental sample of paragraphs and the collection and processing of such data is quite costly. Thus, the fewer the number of paragraphs required the less the cost. This conflict was deemed sufficiently important to justify a more elaborate experiment than the four described in Sections 10 to 13.

Three developmental samples were generated, one being twice the size of the other two. The first consisted of the 666 paragraphs obtained by eliminating every third paragraph from the entire sample of 998 paragraphs. The

TABLE VI

TEST OF NUMBER OF PARAGRAPH CRITERION

(a) Four or More Paragraphs

Assigned \ Actual	U	C	S	Total
U	72	35	21	128
C	31	40	20	91
S	28	27	58	113
Total	131	102	99	332

$$\% \text{ Hits} = \frac{170}{332} = 51.2$$

(b) Ten or More Paragraphs

Assigned \ Actual	U	C	S	Total
U	79	36	26	141
C	26	41	23	90
S	26	25	50	101
Total	131	102	99	332

$$\% \text{ Hits} = \frac{170}{332} = 51.2$$

(c) Fifteen or More Paragraphs

Assigned \ Actual	U	C	S	Total
U	80	36	29	145
C	26	39	21	86
S	25	27	49	101
Total	131	102	99	332

$$\% \text{ Hits} = \frac{168}{332} = 50.6$$

other two of 333 paragraphs each consisted of every other paragraph of the first developmental sample.

The 66 predictor variables described in Section 4 were computed for each sample separately. The screening multiple regression technique, described in Section 3.6 was applied to each group of predictors, and a set of three regression equations was obtained from each of the three samples.

The sets were compared by applying them to the independent sample of data. Application of a set results in three numbers for each paragraph, the highest number determining the assignment of a security classification to that paragraph. The assignments were verified by comparison with the actual classifications. The results, presented in Table VII, show that the percentage of hits of the larger sample (51.2) is higher than that obtained from either of the two smaller samples (50.9 and 46.6, respectively). Therefore, it was concluded that 333 paragraphs were insufficient for a developmental sample.

15. Shrinkage

While conducting the previous experiments, it was noted that the accuracy of security assignments for paragraphs in the independent sample were considerably less than the accuracy for paragraphs of the developmental sample. Such loss in accuracy is termed "shrinkage." In experiment 2, described in Section 11, the percentage of hits fell from 90.3 for the developmental sample to 51.2 for the independent sample. To pursue this point further, the set of three regression equations developed on the large sample of paragraphs (see Section 5.5) was applied to this same sample. As before, the highest of the three numbers obtained for each paragraph determined the assignment of a security classification to that paragraph. Comparison of assigned-versus-actual

TABLE VII

EFFECT OF SAMPLE SIZE IN ACCURACY

(a) Large Sample

Assigned \ Actual	U	C	S	Total
U	64	26	11	101
C	39	42	24	105
S	28	34	64	126
Total	131	102	99	332

$$\% \text{ Hits} = \frac{170}{332} = 51.2$$

(b) First Small Sample

Assigned \ Actual	U	C	S	Total
U	65	26	8	99
C	33	45	33	111
S	33	29	58	120
Total	131	100	99	330*

$$\% \text{ Hits} = \frac{168}{330} = 50.9$$

(c) Second Small Sample

Assigned \ Actual	U	C	S	Total
U	52	30	17	99
C	36	44	24	104
S	42	27	58	127
Total	130	101	99	330*

$$\% \text{ Hits} = \frac{154}{330} = 46.4$$

*Two cases were inadvertently lost. However, this does not materially influence the conclusion.

classifications in given in Table VIII(a). Table VIII(b) is a repeat of Table VII(a) which gives the verification results of applying this same set of equations to the independent set of paragraphs. The percentage of hits drops from 92.9 on the developmental sample to 51.2 on the independent sample. Although shrinkage is likely to occur in any similar statistical analysis, the amount of shrinkage observed here is much greater than usual.

The cause of the shrinkage lies in the method employed in forming the predictor variables. An example serves to illustrate the difficulty. Consider the values for paragraph i of MU100, MC100, and MS100 and assume that paragraph i is secret. The explanation which follows would hold for any paragraph of any security classification and any set of three variables. The values $MU100_i$, $MC100_i$, and $MS100_i$ were obtained by summing the scores of all usable words in paragraph i. The scores were computed by equations (IV-3), repeated below,

$$S_{ju} = (N_{ju} - E_{ju})/(E_{ju})^{1/2}$$

$$S_{jc} = (N_{jc} - E_{jc})/(E_{jc})^{1/2} \qquad \text{(IV-3)}$$

$$S_{js} = (N_{js} - E_{js})/(E_{js})^{1/2} \quad ,$$

where N_{js} is the number of secret paragraphs in which word j appears. Now (the crucial point) if word j appears in paragraph i, a secret paragraph, then N_{js} would tend to be larger than either N_{ju} or N_{jc} simply because paragraph i contributes to N_{js} but not to either of the other two. Therefore, S_{js} will tend to be larger than either S_{ju} or S_{jc}. In fact, <u>all</u> S_{js} values

TABLE VIII

VERIFICATION OF LARGE-SAMPLE REGRESSION EQUATIONS

(a) Developmental Sample

Assigned \ Actual	U	C	S	Total
U	195	4	6	205
C	8	214	6	228
S	7	16	207	230
Total	210	234	219	663

$$\% \text{ Hits} = \frac{616}{663} = 92.9$$

(b) Independent Sample

Assigned \ Actual	U	C	S	Total
U	64	26	11	101
C	39	42	24	105
S	28	34	64	126
Total	131	102	99	332

$$\% \text{ Hits} = \frac{170}{332} = 51.2$$

entering into the formation of $MS100_i$ will tend to be larger than the corresponding values entering into the formation of either $MU100_i$ or $MC100_i$. Therefore, it is not surprising that $MS100_i$ tends to be the largest of the three values. As illustrated by experiment 2 (Section 11), this tendency is strong enough to achieve a percentage of hits of 90 on the developmental sample. This means that in 90 percent of the paragraphs MA100 is larger than the other two predictors, where A is the security classification of the paragraph (A = U, C, or S).

A method was devised to reduce the shrinkage. The developmental sample of data was divided into two equal samples (designated as A and B) by taking every other paragraph. Equations (III-5) were applied to each sample separately to compute word-scores. Thus, the same word has six scores attached to it, three from sample A and three from sample B. Predictor variables were formed as before (Section 9) by summing scores. However, for paragraphs in sample B, word-scores from sample A are applied and vice-versa. The end result is a set of 66 predictor variables for every paragraph in the developmental sample. The screening multiple regression technique was applied to these predictor variables to obtain three regression equations.

To measure the shrinkage, the equations were applied first to samples A and B, which together constitute the developmental sample, and then to the independent sample. To apply the equations to the independent sample of paragraphs, the word-scores developed on A and those developed on B were averaged. That is, a word appearing in a paragraph of the independent sample first had six scores attached to it. These were reduced to three by averaging the corresponding sample A and sample B scores. After averaging, the predictor variables were formed as before by summing, or counting, over usable words of a paragraph.

As in Section 15, the highest of three resulting values for a paragraph determines the assignment of a security classification to that paragraph. The assignments were verified by comparison with the actual classifications and the results are given in Table IX(a) for the developmental sample and Table IX(b) for the independent sample. The percentages of hits are 52.9 and 51.1 for the developmental and independent sample, respectively.

Although the method was quite successful in reducing shrinkage, it did not succeed in raising the absolute level of the accuracy of security assignments to the paragraphs in the independent sample. This can be seen by comparing Table VIII(b) with Table IX(b). The first measures accuracy on the independent sample using the usual type predictors, whereas the second measures accuracy of the "shrink-resistant" predictors. The percentages of hits are almost the same, and the tables are quite similar. This was quite a disappointment to us. Nevertheless, the method is considered to possess considerable merit because it leads to much more realistic estimates of the accuracy to be expected on an independent sample of data. In many problems, independent data may not be available or may be too costly to collect and process.

TABLE IX

VERIFICATION OF SHRINKAGE RESISTANT REGRESSION EQUATIONS

(a) Developmental Sample

Assigned \ Actual	U	C	S	Total
U	90	53	31	174
C	61	110	38	209
S	55	72	149	276
Total	206	235	218	659*

(b) Independent Sample

Assigned \ Actual	U	C	S	Total
U	59	28	14	101
C	31	40	16	87
S	40	32	69	141
Total	130	100	99	329*

*A few cases are omitted because of the method of computation.

16. Application of the REEP Technique - Simple Dummying

The regression estimation of event probabilities (REEP) statistical technique was used to develop methods for assigning security classifications. It will be recalled (Section 6) that the first step in the REEP procedure is to generate a set of dummy predictor variables from each of the continuous type predictor variables. Such generation can be done in a number of ways. Two were attempted; one is reported upon in this section and the other in the next section.

As a result of the previously discussed experiments, the following rules were used to form 66 continuous predictor variables:

(a) The entire set of 998 paragraphs was divided into a developmental sample of 666 paragraphs and an independent sample of 332 paragraphs by extracting every third paragraph to form the independent sample.

(b) Equations (IV-3) were applied to the entire developmental sample of paragraphs to compute three scores for each non-function word and word-pair which appeared in at least four paragraphs.

(c) The method of forming predictor variables was as described in Section 9--i.e., scores of usable words or word-pairs were summed or counted and only the first appearance of a word or word-pair in a paragraph was used.

In this first REEP experiment, just one dummy predictor variable was generated from each of the 66 predictor variables. The method of generation was the same in all cases: a dummy predictor variable takes on a value of <u>one</u> if the value of the original variable is greater than its mean and is <u>zero</u> if the value

is equal or smaller than its mean. The REEP procedure was applied to the resulting 66 dummy predictor variables.

As indicated in Section 6, the end-product of such an application is a method for computing the probabilities that a paragraph is unclassified, confidential, or secret. The largest of these three probabilities determined the assignment of a security classification to a paragraph. Assignments were made in this way to all paragraphs in the independent sample. The assignments were verified as before by comparison with actual classifications. The results are given in Table X.

TABLE X

ACCURACY OF ASSIGNMENT BY REEP TECHNIQUE

INDEPENDENT SAMPLE OF PARAGRAPHS

		Actual			
		U	C	S	Total
Assigned	U	67	35	19	121
	C	36	41	20	97
	S	28	26	60	114
Total		131	102	99	332

The percentage of hits is only 50.6%. This is lower than the percentage achieved by the regression equations of Section 15 [see Table VIII(b)].It is not even as good as the accuracy attained by the very simple procedure described in Section 11 (see Table IV).

The cause of such low accuracy was hypothesized to be the rather simple method we employed to generate the dummy predictor variables. Therefore, another method was devised and is reported upon in the next section.

17. Application of REEP Technique - Final System

The second method for generating dummy predictor variables is to find K-1 numbers which separate an original predictor variable into K groups so that prespecified percentages of the total number of cases fall into each group. The allowable percentages for K = 5 and K = 6 are listed in Table XI. The percentages were obtained from Cox [2] and Bryan and Southan [1] who have devised a method for dividing a continuous variable into K groups such that the grouping error is minimized for a stated K.

TABLE XI

PERCENTAGE FREQUENCY DISTRIBUTION OF OBSERVATIONS IN K DUMMIES

K	Percentage of Total Observations					
5	10.9	23.7	30.8	23.7	10.9	
6	7.4	18.1	24.5	24.5	18.1	7.4

Eleven dummy predictor variables were generated from an original variable. This was accomplished by ranking the 666 values of a variable in numerical order from lowest to highest value and counting up to the required percentage of observations. For example, for K = 5 the first number found, call it $L_1^{(5)}$, is the value of the (.109) x (666) = 73rd ranked observation of the variable; the second number, $L_2^{(5)}$, is the value of the (.109 + .237) x (666) = 230th observation; and so on for $L_3^{(5)}$ and $L_4^{(5)}$. Once the four $L^{(5)}$-values are obtained, they are used to generate 5 dummy predictor variables where

Dummy 1 = 1 if $X \leq L_1^{(5)}$; otherwise dummy 1 = 0

Dummy 2 = 1 if $L_1^{(5)} < X \leq L_2^{(5)}$; otherwise dummy 2 = 0

Dummy 3 = 1 if $L_2^{(5)} < X \leq L_3^{(5)}$; otherwise dummy 3 = 0

Dummy 4 = 1 if $L_3^{(5)} < X \leq L_4^{(5)}$; otherwise dummy 4 = 0

Dummy 5 = 1 if $L_4^{(5)} < X \leq X$; otherwise dummy 5 = 0

This same procedure was used to generate another set of six dummy variables for each predictor by using the percentages listed in the last row of Table XI. These are designated as dummy 6, dummy 7, ..., dummy 11.

The entire set of 66 variables was not used to generate dummy variables. This would have resulted in 11 x 66 = 726 dummy predictor variables, and the REEP computer program cannot handle this many variables. It is our judgment, however, that very little is lost by this reduction because the predictor variables are very highly correlated. A 66 x 66 matrix of correlation coefficients was computed between each variable and every other variable. The correlations were quite high; and from inspection of this matrix the 30 variables listed in Table XII were chosen to generate dummy predictor variables.

TABLE XII

VARIABLES USED TO GENERATE DUMMY PREDICTOR VARIABLES*

MU100	FU66	HU1
MC100	FC66	HC1
MS100	FS66	HS1
MU66	FU50	HU3
MC66	FC50	HC3
MS66	FS50	HS3
MU50	FU34	HU5
MC50	FC34	HC5
MS50	FS34	HS5
MU34		
MC34		
MS34		

*See Section 4 for definitions of the variables.

There were 11 x 30 = 330 dummy predictor variables.

The REEP technique was applied. The dummy predictor variables selected by the technique and the equation generated are shown in Table XIII.

TABLE XIII

DUMMY PREDICTORS SELECTED BY REEP

	Predictors		Predictand		
	Original	Dummy No	Unclassified	Confidential	Secret
B(0)			0.15546	0.60987	0.23466
B(1)	MU100	11	0.38873	-0.33865	-0.05007
B(2)	MU100	12	0.35807	-0.28741	-0.07066
B(3)	MS100	11	-0.08246	-0.17737	0.25983
B(4)	MS100	12	-0.08028	-0.27587	0.35615
B(5)	MS100	4	-0.09212	-0.25260	0.34473
B(6)	MU100	10	0.36366	-0.22026	-0.14340
B(7)	MS100	4	-0.08509	-0.30858	0.39367
B(8)	MU100	5	0.46329	-0.29042	-0.17288
B(9)	MC100	4	-0.15311	0.38555	-0.23243
B(10)	MC100	5	-0.17295	0.40656	-0.23361
B(11)	FS50	4	-0.01907	-0.10652	0.12559
B(12)	MC100	8	0.06903	-0.10464	0.03561
B(13)	MU100	4	0.08837	-0.11252	0.02415

Inspection of Table XIII indicates that 12 dummy predictors were selected, and that all but one of them originated from predictor variables formed by summing all usable words of a paragraph. This has the quite interesting implication that it is not simply the appearance of one or two strong words which governed the original classification of the paragraphs but rather the totality of the words.

The regression equations listed in Table XIII appear to be quite reasonable. Consider the equation for the unclassified predictand. The first predictor selected is MU100, dummy 11 and its coefficient is +0.38873. This variable takes on a value of one if MU100 is in the next to highest of six categories (i.e., $L_4^{(6)} < \text{MU100} \leq L_5^{(6)}$). Therefore, the probability that a paragraph is unclassified

is increased by 0.30113 if MU100 is high. Further inspection reveals that the probability of a paragraph being unclassified decreases is either MS100 or MC100 is high. This too is quite reasonable.

The three equations listed in Table XIII were applied to each paragraph of the independent sample, and the highest of the three resulting values for a paragraph determined the assignment of a security classification to that paragraph. The verification results are shown in Table XIV. The percentage of hits was 53.9 which is higher than any of the percentages achieved heretofore. This agreed with our expectations since our experience in previous investigations of a similar nature has indicated that REEP is the most logical and natural technique to use.

TABLE XIV

VERIFICATION OF REEP EQUATIONS ON INDEPENDENT SAMPLE

		Actual			
		U	C	S	Total
Assigned	U	61	22	9	92
	C	43	59	31	133
	S	27	21	59	107
Total		131	102	99	332

Several paragraphs misclassified by the Final REEP system were examined subjectively in an attempt to determine causes of error. Ten paragraphs assigned a high probability of being secret but which actually were unclassified, and five paragraphs assigned unclassified but actually secret were presented to two scientists with long experience in chemical-biological warfare problems. Both scientists in the past have been charged with assigning security classifications.

Independently, both scientists could see nothing in any of the five actually secret paragraphs which they felt would cause the paragraphs to be secret. For the ten actually unclassified paragraphs, the scientists agreed that they should be unclassified. Thus, the two scientists agreed in one case and disagreed in another with the original classifiers. The examination did not reveal any obvious means of improving the REEP system.

SECTION VI

CONCLUSIONS AND RECOMMENDATIONS

An automatic computer method was devised for assigning a security classification--unclassified, confidential, secret--to a paragraph. The assignment depends entirely on scores attached to the words of the paragraph. The method was tested on an independent sample of 332 paragraphs, and 53.9 percent were correctly classified. Application of the binomial test indicates that 53.9 percent is statistically significantly greater than the 33.3 percent which could be achieved by chance. Therefore, it is concluded that the method does show skill.

Although the automatic method is better than chance, it is not known how it would fare when compared with the present subjective method of assigning security classifications. It is our opinion that the skill is too low to consider replacing the present subjective method of classification. To test this opinion, it is recommended that a study be conducted to measure the skill of the subjective forecaster. The simplest way to do this would be to have two persons independently classify the same set of paragraphs. The amount of matching of the two classifications would be a measure of skill. If this amount exceeds 53.9 percent then the subjective classification is better than the automatic method. Use of more than two classifiers would enhance the confidence in the results. In either case, such an experiment would not only provide a benchmark for measuring the success of objective methods, but it would provide valuable insight into present methods of security classification.

Even if the objective method turned out to be not as good as the subjective method, it might prove quite useful in providing guidance to a classifier. To this end, it is recommended that two experiments be conducted. The first is

rather simple whereas the second is more complicated and costly but offers greater potential usefulness.

The first experiment entails use of the probabilities produced by the REEP technique. It will be recalled that the REEP procedure results in the probabilities that a paragraph belongs to each of the three security classifications. The experiment consists of supplying such probabilities to a number of classifiers for a period of time and then surveying them to determine their opinion of the usefulness of the probabilities. An alternative is to have a group of classifiers assign two security classifications to a set of paragraphs--the first without using the probabilities and the second after seeing the probabilities. Appropriate verification procedures could then determine whether or not the probabilities are useful.

The second, and more elaborate, experiment involves a "sterile environment" classification. It is generally agreed that there is a tendency to overclassify documents since there is no penalty for overclassification as there is for underclassification. A group of classifiers would be asked to assign classifications anonymously to the same sample of paragraphs. The consensus would be the "correct" classification. The objective technique would be developed on this sample of paragraphs. Now, the objective technique would be applied to a new set of paragraphs and the resulting probabilities would be provided to a classifier as guidance. The worth of the guidance information would be evaluated as before. Hopefully, this procedure would reduce the amount of overclassification. A by-product of such an experiment would be a comparison of each classifier against the consensus to determine the number of matches, i.e., the first experiment recommended would be a by-product of this experiment.

It is our firm opinion that the methodology developed and applied during the course of the study has extracted substantially all of the information contained in word and word-pair frequencies. It is recommended that subsequent security classification investigations concentrate on other types of information--e.g., meanings of words, context surrounding the paragraph, rules used to assign security classifications, etc.

It is strongly recommended that the methods be applied to develop objective methods for indexing textual material. From a strictly methodological viewpoint, the three security classifications could have been any three categories--history, biology, mathematics or number theory, calculus, topology--and the procedures would have been the same. Of particular interest in the methodology are (a) the method of scoring words and combining the scores, and (b) the flexibility of the REEP statistical technique. As far as is known, neither has been applied before in automatic indexing procedures. The usual method is to choose "key" words to do the indexing. The proposed methodology would permit both key words and word scores to be presented to the REEP technique in a large variety of ways for objective selection and optimum organization of useful information into an automatic-computer indexing system.

APPENDIX

A.0 ITEMS FORWARDED TO SPONSORING AGENCY

Under terms of the contract, certain items were furnished to the Sponsoring Agency. The items are described briefly below.

A.1 Data

A.1.1 Cards

998 paragraphs were punched onto IBM cards following the rules given in Section 2. The procedures of Section 4. were followed to edit the card data and insert corrected cards when necessary. The corrected data consist of some 13,000 cards. There are two types of cards, a header card at the beginning of each paragraph followed by the data cards for that paragraph. The format of the header card is:

Columns	Description
1-2	$$
3	Security classification of the paragraph; U = unclassified, C = confidential, S = secret.
4-6	Paragraph number, 1,2,...,998. Paragraphs were numbered consecutively as they were extracted from the documents listed in Section 2.6.

Punching of the data cards starts in column 1 and continues through column 70 onto column 1 of the next card through column 70, etc. One word can overlap two cards. Each paragraph starts on a new card. Columns 71-80 are left blank.

A.1.2 Raw Data Tape

The cards were processed by the RAW DATA TAPE GENERATOR PROGRAM described in Section A.2.2 to produce the Raw Data Tape. The tape is unlabelled and was written on an IBM 360 using logical IOCS. The tape is 9-track, 800 bytes per

inch, fixed format of 32 bytes per record and 50 records per block. The record format follows:

Columns	Format	Description
1-8 9-16	Binary Coded Decimal (BCD) BCD	Primary Word) Secondary Word) The first and second words, respectively, of a word-pair. See Section 2.3.2 for a definition of word-pair. Words at the end of a sentence are not paired so the secondary word is left blank. Only the first six characters of a word are used and these are left-adjusted.
17-20	Integer (I)	Paragraph Number
21-24	I	Sentence number within paragraph.
25-28	I	Number of the IBM card (see Section A.1.1) on which primary word begins.
29-32	I	Classification of paragraph; 1 = unclassified, 2 = confidential, 3 = secret.

There is one record for each non-function word appearing in the entire sample of 998 paragraphs. (See Section 3.3 for definition of non-function words.) A word appearing N times will generate N records. The order of the records is by word within paragraph, i.e., first word of first paragraph, second word of first paragraph, ..., last word of last paragraph. There are 65,352 records.

The tape is classified SECRET.

A.1.3 Basic Data Tapes

There are two Basic Data Tapes, one for the 666 paragraphs of the dependent or developmental sample and the second for the 332 paragraphs of the independent or test sample. Both tapes are unlabelled and written on an IBM 360 using logical

IOCS. The tapes are 9-track, 800 bytes per inch, fixed formats of 44 bytes per record and 30 records per block. The record format follows:

Columns	Format	Description
1-4	Integer (I)	Primary word number*
5-8	I	Secondary word number*
9-12	I	Number of times word or word-pair appears in paragraph.
13-16	Floating Point (F)	S_{ju} Unclassified score attached to word or word-pair
17-20	F	S_{jc} Classified score
21-24	F	S_{js} Secret score (see equations (4-3), Section 4.1 for definitions of scores).
25-28	I	N_{ju} The number of different unclassified paragraphs in which the word or word-pair appears at least once.
29-32	I	N_{jc} Same as N_{ju} but for confidential paragraphs.
33-36	I	N_{js} Same as N_{ju} but for secret paragraphs.
37-40	I	Paragraph number
41-44	I	Paragraph classification; 1 = unclassified, 2 = confidential, 3 = secret.

*The actual words are not used. Instead, numbers are assigned to each primary word and secondary word. This allows a security classification of UNCLASSIFIED which provides access to the computer at all times.

There is one record for each usable word in a paragraph (see Section 4.2.1 for a definition of usable). There are 32,548 records on the developmental sample Basic Data Tape and 15,485 on the independent sample tape. Both tapes are UNCLASSIFIED.

A.2 Computer Programs

A.2.1 Statistical Techniques

Computer programs for applying the two statistical techniques and for forming dummy predictor variables were written prior to this contract. (See Section 3 for a description of the techniques and of the dummying procedure.) The programs are on the TRC Statistical Program Tape and are described in a manual prepared for the U.S. Air Force [6]. A copy of the tape and five copies of the manual were furnished to the Sponsoring Agency.

A.2.2 Raw Data Tape Generator

The objective of the program is to generate the Raw Data Tape described in Section A.1.2. This is accomplished by processing the cards described in Section A.1.1. Each paragraph is treated as one long serial string of characters. A word is isolated as a successive group of characters preceded and followed by either a blank or end-of-sentence mark. The word extremes, beginning and end, are then stripped of non-permissible characters. The stripped word is set to a maximum length of six characters by dropping all characters after the first six. Stripped words with less than six characters are left-adjusted and padded with blanks.

A function table search is executed and function words are eliminated. See Section 3.3 for a list of function words. Each non-function word generates one record on the output tape. The format of such records is described in Section A.1.2. Words are isolated and processed in this manner until the set of input cards is exhausted.

The program counts the number of non-function words, the number of function words, and the total number of words in each paragraph. The mean and standard

deviation of each of these three quantities is obtain separately for the unclassified, confidential, and secret paragraphs. These 18 values are printed. Under program option each record placed onto tape can also be printed.

The program is written in assembler language for the IBM 360 Mod. 40. Input-output operations are overlapped for efficiency.

A.2.3 Basic Data Tape Program

The program produces the two tapes described in Section A.1.3. The records on the Raw Data Tape, described in Section A.1.2, have previously been alphabetized first by primary word then by secondary word within primary word. For each word and word-pair, counts are made to obtain N_j, N_{ju}, N_{jc}, N_{js} which are, respectively, the total number of paragraphs in which word j appears at least once and the number of such paragraphs which are unclassified, confidential, and secret. The program also computes the number of times each word or word-pair appears in each paragraph. For those words and word-pairs for which N_j equals or exceeds four, equations (IV-3), Section 8, are applied to compute S_{ju}, S_{jc}, S_{js}. The information is placed onto tape in the format given in Section A.1.3.

In addition to producing the two tapes, the program has three print options which permit it to be used for other purposes.

a. The program can print those words which appear less than M times, where M is an input value, in the entire sample of data. This is useful for locating misspelled words.

b. Each record can be printed as it is placed onto the output tape. It is to be noted that such printing includes the number of times each word appears in each paragraph so that the program is a frequency-count program.

c. The total number of times a word appears in the entire sample of data can also be printed. (This is not the same as N_j, the total number of paragraphs in which word j appears.)

The program is written in assembler language for the IBM 360 Mod. 40. Input-output is overlapped for efficiency

A.2.4 Prediction Program

The output of both the screening multiple regression program and the regression estimation of events probability programs described in Section A.2.1 consists of three equations of the form

$$Y_{ui} = A_{u0} + A_{u1}X_{u1i} + A_{u2}X_{u2i} + \ldots + A_{uP}X_{uPi} \qquad (A\text{-}1)$$

$\hat{Y}_{ui}$ is an estimate of the probability that paragraph i is unclassified. The X's are predictor variables selected by the programs, continuous type predictors by screening regression and dummy predictor variables by REEP. The A's are regression coefficients computed by least squares. Two similar equations estimate, respectively, the probabilities that paragraph i is confidential or secret.

The prediction program applies the three equations to each of a sample of paragraphs. The largest of the three Y_i's determines the security classification assigned to paragraph i. The assignments are matched with the actual classifications and a three by three contingency table is formed and printed. See Section V for several examples of contingency tables.

The program is written in FORTRAN IV for the IBM 360 Mod. 40 computer.

REFERENCES

1. Bryan, J. G., and J. R. Southan, 1962: Optimum subdivision of a variable by the method of D. R. Cox. Scientific Report TRC-21, Contract AF19(604)-5207, The Travelers Research Center, Inc., Hartford, Conn.

2. Cox, D. R., 1957: "Note on grouping." J. Amer. Stat. Assn., Vol. 52, No. 280, pp. 543-7.

3. Enger, I., J. A. Russo, Jr., et al, 1964: A statistical approach to 2-7-hr prediction of ceiling and visibility, Vols. I and II. Technical Reports 7411-118 and 118a, The Travelers Research Center, Inc., Hartford, Conn.

4. Miller, R. G., 1958: Regression estimation of event probabilities. Technical Report 7411-121. Contract Cwb-10704. The Travelers Research Center, Inc., Hartford, Conn.

5. _____, 1962: Statistical prediction by discriminant analysis. Meteorol. Monogr., 4:25, Am. Meteorol. Soc., Boston.

6. Sorenson, E. L., I. Enger, and T. G. Johnson, 1965: User's manual for statistical computer program package. 433L Systems Program Office, Electronics Systems Division, Air Force Systems Command, USAF, L. G. Hanscom Field, Bedford, Massachusetts.

7. Veigas, K. W., 1961: Prediction of twelve, twenty-four, and thirty-six hour displacement of hurricanes by statistical methods. Technical Report 61-3, Contract Cwb-9807. The Travelers Research Center, Inc., Hartford, Conn.

8. Williams, J. H., Jr., 1966: Discriminant analysis for content classification. Technical Report No. RADC-TR-66-6, Information Processing Branch, Rome Air Development Command, Griffiss Air Force Base, New York.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Travelers Research Center, Incorporated 250 Constitution Plaza Hartford, Conn. 0613	Unclassified
	2b. GROUP N/A

3. REPORT TITLE

"Automatic Security Classification Study

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final

5. AUTHOR(S) (First name, middle initial, last name)

Enger, Isadore
Merriman, Guy T.
Bussemey, Ann L.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
October 1967	66	8

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F 30 602-67-C-0042 b. PROJECT NO. 5581 c. 02 d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) RADC-TR-67-472

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Rome Air Development Center (EMIIH) Griffiss AFB, N.Y. 13440

13. ABSTRACT

An investigation was made of the feasibility of using computers to assign the proper security classification (unclassified, confidential), secret) to textual material. The words in 998 paragraphs were transformed to computer-usable form. A set of 66 variables was computed for each paragraph by a two-stage process of attaching three scores to a word and then combining the scores in various ways over the words of a paragraph. Several experiments were conducted to validate assumptions involved in the method of scoring the words and the methods for combining the scores. The 66 variables were presented to a statistical technique which made a preferential selection of a small set of effective variables from the large set of 66 variables. The redundant or non-controlling variables were eliminated from subsequent analysis, and an objective system was developed for assigning security classifications using only the selected variables. The system was applied to an independent sample of paragraphs and 53.9 percent were correctly classified. It was concluded that the system does exhibit skill. However, the skill is probably too low to consider replacing the present system. Finally, it is concluded that the method for forming variables and the statistical technique, both apparently new to this field, show sufficient promise to merit application to other automatic indexing problems.

DD FORM 1473

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Text Processing Discriminant Analysis Automatic Classification						